RESEARCH STATEMENT

Yong Jae Lee, Associate Professor, Computer Sciences Dept. UW-Madison

***Note:** Here's a **link** to a research talk I gave in April 2020, highlighting some of the works in this document.*

My dream is to build machines that can understand the visual world without any human supervision. Humans and animals learn to see the world mostly on their own, without supervision, yet today's state-of-the-art visual recognition systems rely on millions of manually-annotated training images. This reliance on labeled data has become one of the key bottlenecks in creating systems that can attain a human-level understanding of the vast concepts and complexities of our visual world. Indeed, while AI technology is increasingly being used to impact various facets of our daily lives – including commerce, healthcare, transportation, agriculture, and security – most real-world applications are limited to specific domains in which lots of carefully-labeled data can be unambiguously and easily acquired.

To address this limitation, ***my research in computer vision and machine learning strives to create scalable recognition systems that can learn to understand visual data with minimal human supervision.*** In particular, my current research focuses on two main themes: (1) learning to see with weak or no human supervision, and (2) learning to see using video. These themes are *two sides of the same coin*: both are needed for creating systems that can learn to see with minimal human supervision. Below, I first elaborate on my key contributions along these two themes, and then briefly describe computer graphics and cross-disciplinary applications. I will conclude with ongoing and future directions.

Research Progress

1 Learning to See with Weak or No Human Supervision

Low-cost cell phones and cameras, along with social media and photo-sharing websites, have made the Web an endless supply of images and videos; e.g., Facebook reports 350 million photo uploads per day and YouTube sees 500 hours of video uploaded every minute! These images and videos are replete with meta-data such as text tags, GPS coordinates, timestamps, and social media sentiments. The only way to fully take advantage of this huge resource – without any additional annotation effort – requires algorithms that can learn with weak[1] or no human supervision.

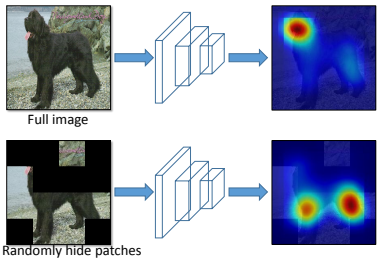
I have created novel *weakly-supervised* algorithms that learn only from weak image-level annotations (e.g., an image of a dog tagged as "dog" without any box or pixel annotations) to detect and segment objects [22, 19, 34, 17, 18] and discover and localize visual patterns that characterize a property of an object [23, 35, 16], as well as *unsupervised* algorithms that learn to discover novel object categories [26, 25, 29, 30, 40] and disentangle latent factors in generative modeling [15, 33, 7, 8]. I summarize each in turn below. This research theme is being supported by my NSF CAREER, NSF EAGER, Adobe Data Science Research Award, and Sony Focused Research Award grants.

Weakly-supervised object detection and segmentation. Detecting and segmenting objects in images is a core problem in computer vision, but today's algorithms require laborious bounding box or pixel-level annotations which are costly and error-prone. For example, to create MS COCO – the de facto benchmark dataset for training and evaluating detection and segmentation algorithms – more than *70,000* hours were spent in annotating 328K images for *only 80* object categories. Clearly, this is not a scalable solution for creating machines that can recognize hundreds of thousands of different visual concepts as we humans do.

I have created novel algorithms for detecting and segmenting objects with only image-level tag annotations [22, 19, 34, 17, 18]. In this setting, the goal is to obtain bounding box or pixel-level classifications of

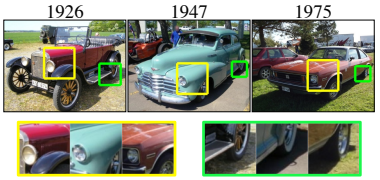
[1] By "weak" supervision, I am referring to the setting in which a method learns with only image-/video-level annotations (e.g. text tags) during training, yet produces more detailed predictions (e.g. bounding box, keypoint, pixel segmentation) during testing.

objects given only image-level labels. I am particularly excited about our recent *Hide-and-Seek* approach [17], which is a surprisingly simple yet highly-effective solution. The key idea is simple: randomly hide image patches in the training images when learning an image classification model. This forces the model to focus on the different (randomly) retained object parts across the training images, which leads to the model learning to localize the entire object (e.g., the entire dog) as opposed to prior methods which focus only on the most discriminative part (e.g., dog's face). This idea has also proven to be useful as a data augmentation technique for training deep networks, improving the state-of-the-art on a variety of tasks including image classification, segmentation, face recognition, and person re-identification [20].



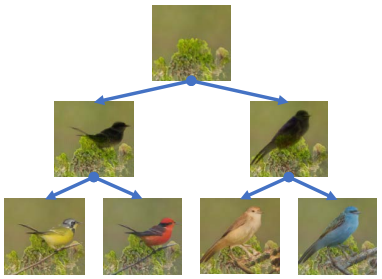
Hide-and-Seek [17] improves weakly-supervised object localization by randomly hiding patches in each training image (bottom), which forces the image classifier to go beyond just the most discriminative part (top) and instead learn to focus on all parts of the object.

Weakly-supervised visual data mining. Apart from scalability, weakly-supervised learning addresses another important issue: for abstract visual concepts like "what makes an antique car look antique?" or "what makes this shoe more comfortable than this other one?", it is often *ambiguous* to know exactly what to label in the image. For example, given an image of an antique car, it can be difficult to precisely annotate at the pixel-level all regions of the car that make it look antique. My research has provided some critical first steps in addressing this difficulty by *automatically discovering* such visual concepts in a data-driven way – by mining patches that are correlated with the weak image labels (e.g., the year that the cars were made; see Figure on right) [23, 35, 16] – as well as by leveraging external knowledge bases and image captions for weakly-supervised object detection and segmentation [34, 14]. For the latter, by leveraging common-sense cues derived from knowledge bases, my algorithms significantly improve upon prior methods that only rely on visual information.



Given historic car images, my algorithm in [23] automatically discovers visual elements (yellow, green boxes) whose appearance variations capture the changes in car style across time.

Unsupervised and self-supervised learning. Weakly-supervised learning does not fully address the scalability issue of visual recognition systems, as it still requires annotations (albeit weak). Ultimately, the holy grail in computer vision is to create recognition systems that can learn without *any* annotations. My research has made fundamental contributions to unsupervised object category learning ("discovery"), in particular with the ideas of *self-paced category discovery* [29] in which objects are learned in order of predicted difficulty, and *context-aware category discovery* [27, 30] in which the growing pool of learned categories serve as context to help discover new unknown categories. These ideas have inspired a large number of work not only in discovery, but also in object detection, image segmentation, and unsupervised representation learning. Finally, I am very excited about our recent works on generative modeling, FineGAN [15] and MixNMatch [7], which are among the first unsupervised methods to yield a structured, disentangled representation of background, object shape, color/texture, and pose for image generation. Building on this work, I recently developed a novel unsupervised generative model for learning disentangled representations in class-imbalanced data [8], which better reflects real-world distributions.

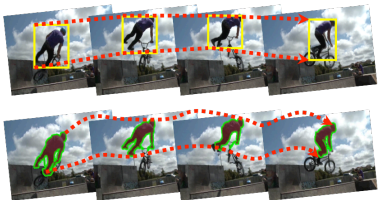

FineGAN [15] is a generative model that learns to hierarchically disentangle the background, object's shape, and object's texture/color for image generation without any mask or object label supervision.

2 Learning to See using Video

Another exciting direction that I have taken is training visual recognition systems with video. Video offers something that images cannot: it provides motion information, which facilitates visual recognition in human vision; e.g. the slithering of a snake or the fluttering of a butterfly helps in their identification. However, research in critical vision tasks such as general (non-human) object classification, detection, and segmentation in *video* have been significantly lagging compared to their image counterparts,

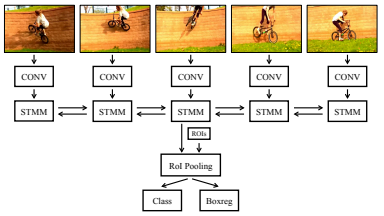
largely due to the huge annotation costs and hardware requirements that video demands. Given our *dynamic* visual world, I contend that these traditional image-based tasks need to be studied with videos, especially since motion is an indispensable cue (*that comes for free!*) for learning to see. Undoubtedly, video will play a critical role in creating machines that learn to see with minimal human supervision.

I have created novel algorithms that segment and detect objects [28, 19, 36, 37] in video as well as algorithms that summarize videos captured from a wearable camera [24, 31, 4]. This research theme is being supported by my Army Research Office Young Investigator Program (ARO YIP), NSF IIS RI Core, and AWS Machine Learning Research Award grants.



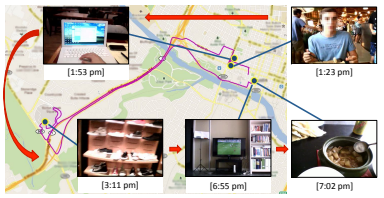
Keysegments [28] and **Track-and-Segment** [36] are unsupervised video object segmentation approaches that automatically identify and segment the foreground objects in unlabeled video.

Video object segmentation and detection. My *Keysegments* work on unsupervised video foreground object segmentation [28] was one of the first to introduce the problem (prior methods required human annotation or segmented out all objects without identifying the foreground ones), and showed how appearance and motion saliency can be used to *discover* prototype instances of the foreground objects. My follow-up *Track-and-Segment* paper [36] showed that self-paced learning can facilitate unsupervised video object segmentation; i.e., by focusing on the easiest frames for initialization, and incrementally updating the segmentation model using new (harder) instances that are discovered and segmented. More recently, I introduced an approach that provides spatio-temporal alignment of the latent memory in recurrent neural networks for supervised video object detection [37]. By aligning the stored visual representation (memory) over time, more accurate spatially-localized visual features can be produced for each object in each video frame. I am currently working towards *unsupervised* video recognition methods that exploit such spatio-temporal alignment.



Spatial-Temporal Memory Networks [37] perform video object detection by learning to model and spatially-align an object's long-term temporal appearance and motion dynamics.

First-person "egocentric" video summarization and analysis. I created the first approach that predicts important objects to summarize hours-long *egocentric* videos captured from a wearable camera (e.g., GoPro) [24, 31]. Egocentric videos offer a first-person view of the world (e.g., we can often see the camera wearer's hands), and can be used to record the daily lives of the user – which is especially valuable for people with memory loss as they provide visual cues to spark back memory. The first-person view also translates naturally to robotics applications and enables a fruitful platform for *embodied* vision research in which agents learn to perceive and act through interaction with their environment. In recent work, together with Indiana University collaborators, I created an algorithm that identifies the first-person camera wearer in a third-person (environmental) video that captures the scene [4]. This work is one of the first to combine information from both first- and third-person videos, which is a setting that is more likely to become common as environmental and wearable cameras become even more ubiquitous. I am excited to continue exploring new questions and solutions in this novel research space.



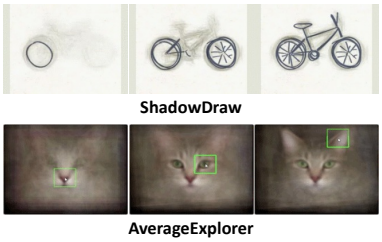
My first-person video summarization algorithm [24, 31] produces keyframe summaries focused on the automatically predicted important people and objects that the camera wearer interacted with.

3 Computer Graphics and Cross-Disciplinary Applications

I enjoy applying my vision and learning algorithms in creative ways. For computer graphics, I have created two novel systems (both published in SIGGRAPH): ShadowDraw [32] and AverageExplorer [41]. ShadowDraw is a real-time interactive system that guides the freeform drawing of objects on a PC tablet – it automatically retrieves and blends images that match the user's ongoing drawing from a large image database.

AverageExplorer is a real-time interactive system that allows a user to rapidly explore and visualize a large image collection using the medium of average images.

Being at UC Davis, I have had the opportunity to collaborate with world-class Veterinary Medicine and Animal Science researchers. I took upon this opportunity to explore two problems: (1) understanding rodent behavior [11], and (2) decoding pain in livestock animals, which involves automatically detecting keypoints (e.g., eyes, nose, mouth) on their faces [10]. The latter is a large project, for which I received the Hellman Fellowship Award, that involves collaborators in the UC Davis Center for Equine Health, Animal Science, Swedish Agricultural University, and UCSD. Together with collaborators in the ECE department, I have also worked on analyzing the adoption and propagation of content (e.g., images) in online social networks [9, 6].



ShadowDraw [32] (top) and **AverageExplorer** [41] (bottom) are real-time interactive systems for freeform drawing and image exploration, respectively.

Ongoing and Future Directions

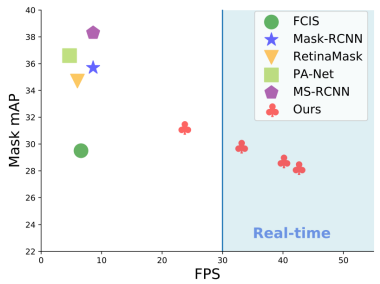
To summarize, my research in computer vision and machine learning focuses on algorithms that learn to understand visual data with weak or no human supervision, and by leveraging motion and temporal cues in video. In addition to these themes, ***I am interested in all other challenges that need to be addressed in creating machines that can attain a human-level understanding of the visual world.*** Specifically, I am interested in exploring questions such as:

- *Can we develop perception algorithms that can learn from multiple modalities?* We humans learn about our world through signals acquired from multiple sources (e.g., sound, vision, smell, touch, taste), which often supervise each other. However, until very recently, computer vision research has largely focused only on utilizing visual data. I believe that multi-modal learning will be especially critical for creating systems that can learn without human annotations. I have begun to make progress in this direction [34, 14, 39, 12, 38].

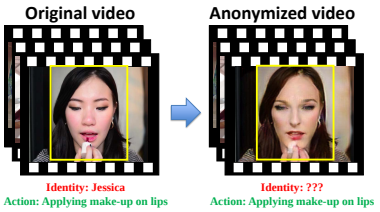
- *Can we create algorithms that can dynamically adapt to changing environments?* While most existing visual scene understanding research assumes a fixed and static environment, this assumption does not hold in many real-world scenarios. Instead, *robust* and *fast* algorithms that can *adapt online* to constantly changing environments are needed. Our recent work on real-time instance segmentation YOLACT [1, 2] takes a step towards this direction.

- *How can we create unbiased and secure visual recognition algorithms?* As computer vision technology is becoming more integrated into our daily lives, addressing ethical, bias/fairness, and privacy/security questions are more important than ever. I have begun to study ways to ensure the privacy and security of users in the visual data that the algorithms process [13, 5, 3], to mitigate undesirable biases [21], and to improve robustness of deep networks [42]. In particular, [42] proposed a novel anti-aliasing module for convolutional networks, and received the **best paper award** at BMVC 2020.

Over the next decades, I will strive to continue to be on the forefront in creating machines that can learn with minimal human supervision. I am passionate about asking the *right* (meaningful and impactful) research questions, and proposing innovative and effective solutions to those questions. I am excited about the prospects of working towards these challenges with collaborators in vision and learning, and related fields including graphics, robotics, neuroscience, and cognitive science.



YOLACT [1] is the first *real-time* (above 30 FPS) instance segmentation approach with competitive instance segmentation accuracy on the challenging MS COCO dataset.



Our **privacy-preserving action detector** [13] learns to modify video frames to anonymize a person's face (so that Jessica is no longer identifiable), while preserving action information (putting on lipstick).

References

[1] D. Bolya, C. Zhou, F. Xiao, and **Yong Jae Lee**. YOLACT: Real-time Instance Segmentation. In *IEEE International Conference on Computer Vision* (ICCV), 2019. (**oral presentation**).

[2] D. Bolya, C. Zhou, F. Xiao, and **Yong Jae Lee**. YOLACT++: Better Real-time Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), 2020.

[3] Z. A. Din, H. Venugopalan, J. Park, A. Li, W. Yin, H. Mai, **Yong Jae Lee**, S. Liu, and S. T. King. Boxer: Preventing Fraud by Scanning Credit Cards. In *USENIX Security Symposium* (USENIX Security), 2020.

[4] C. Fan, J. Lee, M. Xu, K. Singh, **Yong Jae Lee**, D. Crandall, and M. Ryoo. Identifying First-Person Camera Wearers in Third-Person Videos. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017.

[5] X. Gu, W. Luo, M. Ryoo, and **Yong Jae Lee**. Password-conditioned Anonymization and Deanonymization with Face Identity Transformers. In *European Conference on Computer Vision* (ECCV), 2020.

[6] W. Hu, K. Singh, F. Xiao, J. Han, C. Chuah, and **Yong Jae Lee**. Who Will Share My Image? Predicting the Content Diffusion Path in Online Social Networks. In *ACM International Conference on Web Search and Data Mining* (WSDM), 2018.

[7] Y. Li, K. K. Singh, U. Ojha, and **Yong Jae Lee**. MixNMatch: Multifactor Disentanglement and Encoding for Conditional Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2020.

[8] U. Ojha, K. K. Singh, C.-J. Hsieh, and **Yong Jae Lee**. Elastic-InfoGAN: Generative Modeling of Disentangled Representations in Class-Imbalanced Data. In *Neural Information Processing Systems* (NeurIPS), 2020.

[9] M. Rahman, J. Han, **Yong Jae Lee**, and C. Chuah. Analyzing the Adoption and Cascading Process of OSN-Based Gifting Applications: An Empirical Study. *ACM Transactions on the Web* (TWEB), 11(2), 2017.

[10] M. Rashid, X. Gu, and **Yong Jae Lee**. Interspecies Knowledge Transfer for Facial Keypoint Detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017.

[11] Z. Ren, A. Noronha, A. V. Ciernia, and **Yong Jae Lee**. Who Moved My Cheese? Automatic Annotation of Rodent Behaviors with Convolutional Neural Networks. In *IEEE Winter Conference on Applications of Computer Vision* (WACV), 2017.

[12] Z. Ren and **Yong Jae Lee**. Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2018.

[13] Z. Ren, **Yong Jae Lee**, and M. Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *European Conference on Computer Vision* (ECCV), 2018.

[14] K. Singh, S. Divvala, A. Farhadi, and **Yong Jae Lee**. DOCK: Detecting Objects by transferring Common-sense Knowledge. In *European Conference on Computer Vision* (ECCV), 2018.

[15] K. Singh, U. Ojha, and **Yong Jae Lee**. FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019. (**oral presentation**).

[16] K. Singh and **Yong Jae Lee**. End-to-End Localization and Ranking for Relative Attributes. In *European Conference on Computer Vision* (ECCV), 2016.

[17] K. Singh and **Yong Jae Lee**. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. In *IEEE International Conference on Computer Vision* (ICCV), 2017.

[18] K. Singh and **Yong Jae Lee**. You reap what you sow: Using Videos to Generate High Precision Object Proposals for Weakly-supervised Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019.

[19] K. Singh, F. Xiao, and **Yong Jae Lee**. Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016.

[20] K. Singh, H. Yu, A. Sarmasi, G. Pradeep, and **Yong Jae Lee**. Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond. In *arXiv*, 2018.

[21] K. K. Singh, D. Mahajan, K. Grauman, **Yong Jae Lee**, M. Feiszli, and D. Ghadiyaram. Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2020. (**oral presentation**).

[22] H. O. Song, **Yong Jae Lee**, S. Jegelka, and T. Darrell. Weakly-supervised Discovery of Visual Pattern Configurations. In *Neural Information Processing Systems* (NeurIPS), 2014.

[23] **Yong Jae Lee**, A. A. Efros, and M. Hebert. Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time. In *IEEE International Conference on Computer Vision* (ICCV), 2013. (**oral presentation**).

[24] **Yong Jae Lee**, J. Ghosh, and K. Grauman. Discovering Important People and Objects for Egocentric Video Summarization. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2012.

[25] **Yong Jae Lee** and K. Grauman. Foreground Focus: Unsupervised Learning From Partially Matching Images. *International Journal of Computer Vision* (IJCV), 85, 2009.

[26] **Yong Jae Lee** and K. Grauman. Shape Discovery from Unlabeled Image Collections. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2009.

[27] **Yong Jae Lee** and K. Grauman. Object-Graphs for Context-Aware Category Discovery. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2010. (**oral presentation**).

[28] **Yong Jae Lee** and K. Grauman. Key-Segments for Video Object Segmentation. In *IEEE International Conference on Computer Vision* (ICCV), 2011.

[29] **Yong Jae Lee** and K. Grauman. Learning the Easy Things First: Self-Paced Visual Category Discovery. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2011.

[30] **Yong Jae Lee** and K. Grauman. Object-Graphs for Context-Aware Visual Category Discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), 34(2):346–358, 2012.

[31] **Yong Jae Lee** and K. Grauman. Predicting Important Objects for Egocentric Video Summarization. *International Journal of Computer Vision* (IJCV), 114(1):38–55, 2015.

[32] **Yong Jae Lee**, C. L. Zitnick, and M. Cohen. ShadowDraw: Real-Time User Guidance for Freehand Drawing. *ACM Transactions on Graphics* (Proceedings of SIGGRAPH), 30(4), 2011.

[33] F. Xiao, H. Liu, and **Yong Jae Lee**. Identity from here, Pose from there: Self-supervised Disentanglement and Generation of Objects using Unlabeled Videos. In *IEEE International Conference on Computer Vision* (ICCV), 2019.

[34] F. Xiao, L. Sigal, and **Yong Jae Lee**. Weakly-supervised Visual Grounding of Phrases with Linguistic Structures. In *IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), 2017.

[35] F. Xiao and **Yong Jae Lee**. Discovering the Spatial Extent of Relative Attributes. In *IEEE International Conference on Computer Vision* (ICCV), 2015. (**oral presentation**).

[36] F. Xiao and **Yong Jae Lee**. Track and Segment: An Iterative Unsupervised Approach for Video Object Proposals. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016. (**spotlight presentation**).

[37] F. Xiao and **Yong Jae Lee**. Video Object Detection with an Aligned Spatial-Temporal Memory. In *European Conference on Computer Vision* (ECCV), 2018.

[38] F. Xiao, **Yong Jae Lee**, K. Grauman, J. Malik, and C. Feichtenhofer. Audiovisual SlowFast Networks for Video Recognition. In *arXiv*, 2019.

[39] M. Zhou, R. Cheng, **Yong Jae Lee**, and Z. Yu. A Visual Attention Grounding Neural Model for Multimodal Machine Translation. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2018. (**oral presentation**).

[40] T. Zhou, **Yong Jae Lee**, S. Yu, and A. A. Efros. FlowWeb: Joint Image Set Alignment by Weaving Consistent, Pixel-wise Correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2015. (**oral presentation**).

[41] J.-Y. Zhu, **Yong Jae Lee**, and A. A. Efros. AverageExplorer: Interactive Exploration and Alignment of Visual Data Collections. *ACM Transactions on Graphics* (Proceedings of SIGGRAPH), 33(4), 2014.

[42] X. Zou, F. Xiao, Z. Yu, and **Yong Jae Lee**. Delving Deeper into Anti-aliasing in ConvNets. In *British Machine Vision Conference* (BMVC), 2020. (**Best Paper Award**).